Résumé automatique de textes d'opinion

Aurélien Bossard, Michel Généreux

TALN 2009, Senlis, 24–26 juin 2009

Résumé automatique de textes d'opinions

Michel Généreux et Aurélien Bossard
Laboratoire d'Informatique de Paris-Nord
(CNRS UMR 7030 et Université Paris 13)
99, av. J.-B. Clément – 93430 Villetaneuse
prénom.nom@lipn.univ-paris13.fr

Résumé. Le traitement des langues fait face à une demande croissante en matière d'analyse de textes véhiculant des critiques ou des opinions. Nous présentons ici un système de résumé automatique tourné vers l'analyse d'articles postés sur des blogues, où sont exprimées à la fois des informations factuelles et des prises de position sur les faits considérés. Nous montrons qu'une approche classique à base de traits de surface est tout à fait efficace dans ce cadre. Le système est évalué à travers une participation à la campagne d'évaluation internationale TAC (*Text Analysis Conference*) où notre système a réalisé des performances satisfaisantes.

Abstract. There is currently a growing need concerning the analysis of texts expressing opinions or judgements. In this paper, we present a summarization system that is specifically designed to process blog posts, where factual information is mixed with opinions. We show that a classical approach based on surface cues is efficient to summarize this kind of texts. The system is evaluated through a participation to TAC (*Text Analysis Conference*), an international evaluation framework for automatic summarization, in which our system obtained good results.

Mots-clés : résumé automatique, analyse de textes subjectifs, évaluation automatique.

Keywords: automatic summarization, analysis of subjective texts, automatic evaluation.

1 Introduction

Le résumé automatique a connu un fort renouveau ces dernières années et les recherches en ce domaine ont fortement évolué récemment : l'apparition de gros corpus parfois hétérogènes et la généralisation des techniques d'analyse de surface ont à la fois renouvelé les besoins et les approches. Plus récemment encore, avec l'avènement de médias plus interactifs, la nécessité de repérer les citations, les jugements et les opinions s'est avéré de plus en plus crucial. Le but n'est plus seulement de produire une synthèse de l'information contenue dans les textes, il faut en outre dégager des tendances, identifier les opinions exprimées et si possible en faire la synthèse. Cet article décrit des recherches menées dans ce cadre : nous avons développé un système visant à faire une synthèse automatique d'opinions exprimées sur Internet sur un sujet donné, en particulier dans des articles postés sur des blogues (ci-après *blogues*). Le système repose sur une synthèse entre des techniques classiques de production de résumé par extraction de passages pertinents et l'analyse des opinions exprimées dans ces textes. Les blogues comportant généralement une structure syntaxique irrégulière ainsi qu'une grande richesse néologique, un

minimum de modification est nécessaire pour adapter un tel système pour le traitement d'autres types de documents moins complexes.

Pour évaluer notre système, nous avons participé à la campagne TAC 2008, une campagne d'évaluation internationale organisée par le NIST (*National Institute of Standards and Technology*) et tournée vers les systèmes de questions-réponses et de résumé automatique. L'évaluation proposée dans ce cadre mêlait résumé factuel et analyse d'opinion, à partir des sorties de systèmes de questions-réponses. L'enjeu était de produire des synthèses cohérentes à partir de questions en langage naturel — en général, un résumé correspond à plusieurs questions liées (appelées *squishy list*) sur un thème donné (appelé *target*). Pour prendre un exemple, un des thèmes proposé était la personnalité de l'année désignée par le magazine Time pour 2005 ("*Time Magazine 2005 Person of the Year*"). Une des questions liées était la suivante : "*Why did readers support Time's inclusion of Bono for Person of the Year ?*". On voit qu'il s'agit de questions en "pourquoi" (*why*) : contrairement aux questions factuelles (questions dites *factoïdes*, où la réponse est généralement une entité nommée), il n'est pas possible de répondre de façon simple à ces questions en pourquoi. Les systèmes de questions-réponses traditionnels, qui produisent des fragments en guise de réponse ("*snippets*", c'est-à-dire de courtes séquences de texte issues du fonds documentaire et censées contenir une réponse pertinente) ne sont pas encore tout à fait adaptés pour ce type de problème. L'extraction de phrases donnant une idée des informations essentielles contenues dans le fonds documentaire et formant autant que possible un tout cohérent, semble une voie plus prometteuse. Les participants à la campagne *TAC Opinion 2008* disposaient de fragments extraits par les systèmes de questions-réponses afin de les aider à produire une synthèse cohérente. Nous avons utilisé ces extraits et nous avons ainsi pu élaborer un système très performant par rapport aux autres systèmes évalués[1]. Cependant, cette approche rend notre système dépendant des systèmes de questions-réponses.

Dans ce qui suit, nous présentons un rapide état de l'art du domaine. Nous présentons ensuite le système que le LIPN a développé pour le résumé automatique de textes véhiculant une opinion ; ses différents aspects sont détaillés, notamment les traits de surface utilisés, les procédures de choix des phrases extraites et leur ordonnancement. Nous présentons ensuite les adaptations que nous avons dû faire pour répondre aux besoins particuliers de la campagne TAC, les résultats que nous avons obtenus et leurs limites.

2 État de l'art

La génération de résumés reposant sur des principes linguistiques avancés s'est rapidement avérée trop ambitieuse. C'est pourquoi dès les années 1950, la recherche s'est alors tournée vers des techniques plus sommaires visant à extraire des (segments de) phrases pertinentes. L'idée consiste à attribuer à chaque phrase un poids puis à extraire, en fonction d'un taux de compression, les N phrases dont le poids est le plus élevé.

Les techniques à l'œuvre sont assez intuitives et ont été bien décrites dès les années 1960 (Edmundson, 1969) : repérage de mots clés (par rapport au thème abordé), de "signatures" (syntagmes introducteurs typiques de séquences importantes comme *en résumé*, *en conclusion*, *etc*.), prise en compte de la position des phrases dans le documents, *etc*. Un des systèmes actuellement

[1] Un autre système, appelé LIPN2-Opinion, a également obtenu de bonnes performances sans utiliser les fragments fournis par les systèmes de questions-réponses. Ce système est détaillé dans (Bossard *et al.*, 2008).

les plus populaires, le système MEAD[2] mis au point par Radev, repose sur une approche de ce type (Radev *et al.*, 2001). Le système cherche d'abord à identifier les mots les plus saillants dans chaque texte, les *centroïdes*, et à favoriser les phrases essentiellement constituées des centroïdes. D'autres travaux ont apporté des améliorations à partir de ce schéma de base. On notera par exemple le système Neo-cortex développé à l'Université d'Avignon, qui tire partie de plusieurs mesures de sélection de phrases afin de combiner les avantages de différents types de traits (Boudin & Moreno, 2007). Ce système a récemment obtenu de très bons résultats[3] en utilisant l'algorithme MMR (*Maximal Marginal Relevance*) de (Carbonell & Goldstein, 1998).

Il y a eu des tentatives récentes pour réintroduire de la linguistique profonde dans les systèmes de résumés. Il s'agit essentiellement de travaux opérant directement sur l'arbre syntaxique pour essayer de proposer des procédés permettant de comparer les arbres, de les élaguer ou de les fusionner (Gotti *et al.*, 2007), ou encore de méthodes de compression syntaxique basée sur des propriétés linguistiques théoriques et empiriques (Yousfi-Monod, 2007). Nous en restons dans ce qui suit à une analyse de surface et nous n'avons pas recours à la syntaxe mais ces travaux constituent des perspectives intéressantes et naturelles au travail présenté ici. D'autres approches sont possibles, notamment celles fondées sur la représentation des connaissances (Mani, 2004), la segmentation thématique (Farzindar *et al.*, 2004) ou le profil utilisateur (Châar *et al.*, 2004; Crispino & Couto, 2004).

Dans le domaine des opinions, les travaux précédents se sont surtout attardés à leur détection ainsi qu'à la gradation de leur niveau affectif, et ce selon trois niveaux principaux de sous-tâches. La première sous-tâche consiste à distinguer les textes *subjectifs* des textes *objectifs* (Yu & Hatzivassiloglou, 2003). La seconde sous-tâche s'attarde à classer les textes subjectifs en *positifs* or *négatifs* (Turney, 2002). Le troisième niveau de raffinement essaie de déterminer jusqu'à quel point les textes sont positifs ou négatifs (Wilson *et al.*, 2004). L'impulsion donnée par des campagnes telles que *TREC Blog Opinion Task* depuis 2006 est incontestable (Zhang *et al.*, 2007; Dey & Haque, 2008). Il faut reconnaître que notre traitement des opinions pour l'aide à la production de résumé de textes d'opinion ne va pas au-delà de la distinction standard positif versus négatif, et qu'il faut signaler les efforts récents pour réintroduire des approches plus linguistiques et discursives (prise en compte de la modalité, de l'énonciateur) dans ce domaine (Asher *et al.*, 2008).

En ce qui concerne l'évaluation de résumés, soulignons la contribution de (Goulet, 2007) qui va au-delà de la couverture des n-grammes et propose une terminologie adaptée au français.

3 Description du système

L'approche que nous avons développée combine des techniques traditionnellement employées pour le résumé automatique de textes avec des techniques de détection et d'analyse d'opinions.

3.1 Principes généraux

A la différence des systèmes de résumé traditionnels qui prennent en entrée un texte (ou un ensemble de textes dans le cas du résumé multi-documents), notre système repose fondamenta-

[2] `http://www.summarization.com/mead/`

[3] NIST Document Understanding Conference (DUC) 2006

lement sur une requête qui permet de préciser le fait ou l'objet à propos duquel l'utilisateur souhaite obtenir une synthèse. En effet, il n'y a guère de sens à proposer un système produisant des résumés rendant compte des opinions exprimées si l'on n'a pas précisé d'abord la cible recherchée, c'est-à-dire l'événement, la personne ou l'objet à propos duquel on cherche à connaître l'état de l'opinion. De ce fait, le système repose fondamentalement sur deux éléments : 1) une requête comprenant une cible et éventuellement des questions annexes permettant de préciser l'information recherchée, 2) le fonds documentaire qui sert de base à la production de résumé. On remarquera, du fait de la présence d'une requête, le lien assez direct entre ce type de systèmes et les systèmes de questions-réponses.

Notons à ce propos que la campagne TAC 2008 fournissait en outre une liste de fragments (*snippets* ou *fragments*) préalablement produits par un système de questions-réponses (un *snippet* est un court extrait de texte, de taille fixe ou non, mais ne correspondant pas obligatoirement à une séquence linguistique complète, censé donner un élément de réponse par rapport à une requête). On verra l'importance de ces éléments de réponse pour le système dans la section traitant plus spécifiquement de notre participation à la campagne TAC.

Comme la plupart des systèmes participant aux campagnes d'évaluation comme TAC, notre système repose fondamentalement sur un ensemble d'heuristiques. Celles-ci reflètent indirectement les observations d'analystes quant aux facteurs à prendre en compte pour produire un résumé par extraction. Soulignons dès à présent le manque de fondement théorique de ce type d'approches, leur dépendance plus ou moins grande vis-à-vis du domaine ou du type de texte considéré ; toutefois, comme nous l'avons rappelé dans la section précédente, le peu de résultats des approches génératives a contribué au succès des approches à base d'heuristiques (même si les approches génératives semblent reposer sur des fondements linguistiques plus assurés).

3.2 Architecture

Dans cette campagne TAC, le système construit doit traiter en entrée une cible, une ou deux questions en rapport avec la cible, un groupe de documents (des blogues) pouvant contenir les réponses ou éléments essentiels à nos questions, ainsi que les *snippets*, fournis par TAC mais dont l'usage en entrée est optionelle. En sortie, le système pruduit doit fournir un seul résumé pour chaque combinaison cible/questions/documents.

Cette section décrit les différents modules mis au point et enchaînés lors de l'analyse. Comme les documents analysés sont des blogues, il est dans un premier temps souvent nécessaire de les nettoyer pour extraire le texte et exclure tous les éléments qui peuvent parasiter l'analyse.

Nettoyage Chaque blogue est analysé par un ensemble de programmes visant à extraire le texte et à éliminer toutes les parties bruitées et annexes (balises, javascript, *etc.*).

Sélection des documents pertinents par rapport à une requête La phase de sélection des documents pertinents vise à regrouper les documents par rapport à une requête donnée. La liste des documents potentiellement pertinents était fournie directement par les organisateur de TAC mais un tel module doit être intégré à la plate-forme si le système doit s'appuyer directement sur le fonds documentaire.

Calcul des critères de sélection de phrases Une fois que les textes pertinents ont pu être isolés, il est possible de procéder au calcul des traits pertinents pour l'analyse. Nous prenons en compte quatre types de critères principaux :

- la mesure de l'opinion véhiculée (polarité positive ou négative) ;
- la présence de mots centroïdes ;
- la similarité de la phrase visée avec la requête ou un des fragments ;
- la position de la phrase dans le texte.

Nous détaillons de manière beaucoup plus précise l'ensemble de ces éléments dans les sections qui suivent.

Pondération des phrases La sélection des phrases est faite sur la base d'un vecteur de valeurs issu des calculs effectués à l'étape précédente. Pour éviter d'avoir affaire à des phrases trop courtes (phrases qui peuvent parasiter l'analyse car elles correspondent souvent à des titres ou des parenthèses mal reconnues lors du découpage initial), on exclut toutes les phrases dont la longueur est inférieure à un seuil fixé à l'avance (ici nous avons utilisé un seuil de dix mots).

Pour le reste, un score est attribué à chaque trait, en fonction de son intérêt pour la tâche. Ces scores sont attribués *a priori*, sur la base d'expériences faites sur les données d'entraînement, qui sont parfois disponibles en faible quantité ou qui ne sont pas complètement représentatives des données à analyser *in fine*.

Détection de la redondance et production du résumé Les phrases ayant reçu les scores les plus élevés sont enfin sélectionnées pour produire le résumé, en commençant bien évidemment par la phrase dont le score le plus élevé (censé identifier la phrase véhiculant le plus d'information pertinente par rapport à la cible).

Par ailleurs, comme le relève (Radev *et al.*, 2001), ce type d'approche souffre du fait que des phrases redondantes peuvent être intégrées au résumé. Avant d'intégrer une phrase au résumé, on la compare avec les phrases déjà sélectionnées. Si la phrase en cours d'analyse a un taux de similarité supérieur à un seuil prédéfini, elle n'est pas intégrée.

Nous avons détaillé ici les grands principes qui guident la production de résumés. La liste précise des traits considérés, leurs poids respectifs et les techniques utilisées dépendent de l'application visée, ce qui permet d'avoir un système à la fois générique et très facilement adaptable à de nouveaux besoins.

Nous précisons dans la suite de cette section les techniques de reconnaissance de l'opinion véhiculée. Les sections suivantes détailleront les "instanciations" de cette architecture générique réalisées pour la campagne *TAC Opinion Pilot 2008*.

3.3 Calcul de l'opinion véhiculée dans les textes

Afin de produire un résumé rendant compte des opinions exprimées sur un thème donné, il est nécessaire d'identifier les opinions, leur orientation positive et négative et si possible leur intensité. On peut ainsi regrouper les avis exprimés par groupes cohérents et éventuellement calculer l'orientation générale de l'opinion sur un sujet donné.

Dans les expériences que nous présentons ici, nous nous contentons de classer les opinions en deux grandes classes, positives ou négatives. Pour faire cette classification, notre approche est fondée sur l'utilisation d'un classifieur binaire reposant sur un apprentissage à partir de documents représentatifs préalablement annotés. Le classifieur essaie de repérer automatiquement les éléments pertinents pour une prise de décision (texte véhiculant une opinion positive *vs* texte véhiculant une opinion négative). Nous nous appuyons sur un séparateur à vaste marge (SVM[4])

[4] `http://www.cs.waikato.ac.nz/ml/weka/`

binaire entraînée sur des données typiques pour la tâche (en l'occurrence des exemples de requêtes fournies pour la phase de mise au point et annotées manuellement ainsi que le corpus de films de Cornell[5]). Le but de cette étape — outre la reconnaissance des opinions en elles-mêmes — est de permettre ensuite le regroupement des phrases extraites en fonction de l'opinion exprimée et de tenir compte de la proportion d'opinions allant dans un sens ou dans l'autre, afin d'améliorer le rendu et la lisibilité du résumé.

Comme nous le montrons plus loin, les résultats de l'analyse de l'opinion sur les données de TAC ne semblent pas avoir eu d'effet positif sur les résultats, ce qui est évidemment décevant. Nous analysons ces résultats dans la section suivante, mais disons dès maintenant que l'approche au moyen de SVM ne permet pas de classer directement les phrases, car celles-ci comportent trop peu d'éléments caractéristiques. Le choix que nous avons alors dû faire, à savoir de calculer l'opinion directement au niveau du blogue puis de reporter cette analyse au niveau de chaque phrase est sans doute trop grossier, de nombreuses phrases n'exprimant pas d'opinion en tant que tel. Pour illustrer le lien qui existe entre résumé et analyse d'opinions, considérons la requête positive *"What features do people like about Vista ?"*, pour laquelle le système préférera des phrases tirées de textes positifs pour le résumé (e.g. *"One of the quantum leaps Windows Vista will make is the move from raster graphics to high-quality vector graphics."*), et la requête négative *"What features do people dislike about Vista ?"*, pour laquelle le système préférera des phrases tirées de textes négatifs pour le résumé (e.g. *"Buyers may feel hard done-by with this option, and severely out of pocket purchasing the real Vista experience."*).

4 Expérience : campagne TAC Opinion 2008

Nous détaillons dans cette section l'instanciation de notre architecture pour la campagne *TAC Opinion 2008*. Lors de cette campagne, le système présenté par le LIPN a obtenu des performances très homogènes, étant la plupart du temps classé parmi les cinq premiers systèmes et obtenant même le meilleur résultat si l'on fait la moyenne des différents éléments évalués lors de la campagne.

4.1 Description du corpus

Comme nous l'avons déjà dit, la campagne *Opinion Pilot 2008* portait sur des blogues en anglais et était très liée à la campagne de questions-réponses ayant lieu dans le même cadre. Nous avons donné en introduction (section 1) un exemple représentatif. Rappelons que les résumés doivent correspondre à un thème, précisé par une liste de questions (la "*squishy list*").

Le système disposait en outre d'une liste de blogues potentiellement pertinents pour la question et de fragments fournis par les systèmes de questions-réponses. Rappelons ici qu'un fragment ne correspond généralement pas à une séquence linguistique complète et qu'il peut être erroné (ne pas contenir de réponse pertinente par rapport à la question posée).

[5] `http://www.cs.cornell.edu/People/pabo/movie-review-data/`

4.2 Instanciation du système pour TAC : liste des traits pris en compte

Cette section présente l'instanciation du système détaillé dans la section précédente pour TAC.

Valeur d'opinion : Nous entraînons deux classifieurs binaires, le premier pour classer chacun des blogues, le deuxième pour classer les requêtes. L'analyse des requêtes révèle une variation limitée : elles respectent des patrons syntaxico-sémantiques relativement simples que le SVM capte aisément. Quelques exemples typiques ont suffit à entraîner le classifieur. L'analyse des phrases issues des blogues est beaucoup plus difficile. Celles-ci sont en effet très variables quant à leur longueur, leur structure et leur contenu. Ces observations ont conduit à une approche prudente mais relativement grossière : chaque texte issu d'un blogue a été pris comme un tout et a servi de base à l'apprentissage. La valeur obtenue pour le texte entier a ensuite été affectée à chaque phrase isolée. Cette approche a été adoptée après avoir observé que les blogues étaient très peu souvent nuancés. Ils véhiculent généralement des opinions quasi-uniformément positives ou négatives, ce qui nous a poussé à adopter cette approche où la tendance générale est répercutée sur chaque phrase particulière. Les phrases censées véhiculer une opinion positive reçoivent un score de +1, les phrases véhiculant une opinion négative un score de -1.

Phrase la plus longue : La phrase la plus longue d'un texte reçoit un score de 1, toutes les autres phrases ont un score équivalent à 0. D'autres mesures plus fines sont possibles, comme par exemple le nombre de mots signifiants.

Similarité avec la cible : La similarité de la phrase avec la cible permet d'attribuer un score compris entre 0 et 1.

Similarité avec la requête : La similarité de la phrase avec la requête reçoit de la même manière une valeur située entre 0 et 1.

Similarité avec la première phrase du blogue : La similarité de chaque phrase du blogue considéré avec la première phrase du texte reçoit une valeur comprise entre 0 et 1. Ce calcul repose sur le fait que la première phrase comporte souvent les éléments essentiels, au moins le thème abordé, comme dans le cas des dépêches d'agence.

Similarité avec les fragments : Une liste de fragments est fournie pour chaque cible considérée. Chaque phrase reçoit un score compris entre 0 et 1, correspondant à la valeur la plus grande obtenue lors du calcul (valeur correspondant au trait de similarité entre la phrase et le fragment qui lui est le plus semblable).

Centroïdes : Un score est attribué à chaque phrase *s* en fonction du nombre de mots significatifs (les centroïdes) qu'elle contient. Cette valeur est calculée comme suit :

$$centroid_s = \sum_{mots} term\_frequency_{mot,s} * inverse\_document\_frequency_{mot,s} \quad (1)$$

Cette valeur est finalement normalisée afin d'obtenir un score entre 0 et 1. Pour ce faire, on divise le score obtenu pour chaque phrase par le score obtenu pour la phrase dont le score est le plus élevé.

Position dans le document : Cette valeur reflète la position de la phrase *s* par rapport au début du texte. Elle est calculée comme suit (`sno` = le numéro de la phrase) :

$$position_s = \sqrt{1/sno_s} \quad (2)$$

Les scores attribués à chaque élément sont en outre pondérés d'après la formule 3.

$$\sum_{i=0}^{n} w_i * f_i \tag{3}$$

où w_i représente le poids du trait f_i. Après une phase expérimentale jaugeant la qualité des résumés produits aux vues de différentes combinaisons de poids, la sélection finale des différents poids est présentée dans le tableau 1.

Critère	**Requête positive**	**Requête négative**
Sim...Fragments	+100	+100
Sim...Requête	+40	+40
Sim...Cible	+20	+20
Opinion	+20	-20
Sim...PremièrePhrase	+10	+10
Centroïdes	+10	+10
Position	+10	+10
PhraseLaPlusLongue	-10	-10

TAB. 1 – Poids accordé aux différents traits

On voit ici qu'un poids important a été attribué à la similarité entre les phrases et les fragments (*Sim...Fragments*), ce qui signifie que le système s'appuie assez fortement sur les résultats des systèmes de questions-réponses. Les phrases longues sont légèrement pénalisées.

4.3 Résultats obtenus lors de la campagne TAC

Les résultats sont évalués en utilisant la métrique PYRAMIDE (Nenkova *et al.*, 2007), qui vérifie que les éléments d'informations essentiels (tels qu'on les trouve dans des résumés de référence réalisés par des humains) sont présents. Cette métrique est complétée par une série d'appréciations (notées sur dix par des experts humains) visant à déterminer la qualité du résumé (en termes de lisibilité, grammaticalité, cohérence, fluidité et pertinence). Comme on peut le voir dans le tableau 2, le système *LIPN1-opinion* a obtenu des résultats tout à fait compétitifs par rapport aux autres systèmes (36 *runs* ont été soumis par l'ensemble des participants).

Pyramide	**Grammaticalité**
F-mesure : 0.393	score : 6.636
(premier : 0.534, dernier : 0.101)	(premier : 7.545, dernier : 3.545)
Absence de redondance	**Cohérence**
score : 6.818	score : 3.045
(premier : 8.045, dernier : 4.364)	(premier : 3.591, dernier : 2.000)
Fluidité	**Pertinence**
score : 4.591	score : 4.500
(premier : 5.318, dernier : 2.636)	(premier : 5.773, dernier : 1.682)

TAB. 2 – Résultats de l'évaluation

Si l'on donne le même poids à chacun des six éléments évalués lors de TAC, LIPN1-opinion se classe premier avec un score moyen de 0,492 (dernier : 0,290). Si on regroupe les résultats suivant les trois "axes d'évaluation" proposés par TAC[6] et si on prend en compte aussi bien les résumés produits automatiquement que les résumés "de contrôle", produits par des humains à titre de référence et de comparaison, on obtient les résultats suivants :

Contenu : LIPN1-opinion (F-mesure = 0,393 ; premier : 0,534 – dernier : 0,101) est classé 5^e, derrière un résumé produit à la main et trois produits automatiquement ;

Lisibilité : LIPN1-opinion (score = 4,218 ; premier : 4,873 – dernier : 2,727) est classé 4^e, derrière un résumé produit à la main et deux produits automatiquement ;

Pertinence globale : LIPN1-opinion (score = 4.500 ; premier : 5,318 – dernier : 2,636) est classé 8^e, derrière un résumé produit à la main et six produits automatiquement.

5 Conclusion

Dans cet article, nous avons montré un système permettant de produire automatiquement des résumés de textes porteurs d'opinions, des blogues notamment. Le système repose sur des techniques classiques en résumé : le calcul de différents traits reposant sur des heuristiques découlant de diverses expériences sont à la base de notre approche.

Etant donné les résultats que nous avons obtenus pour la campagne TAC 2008, nous pensons que ce système, encore très basique, constitue néanmoins un bon point d'entrée pour cette tâche. Nous avons aussi montré qu'au-delà des traits spécifiques mis en place pour TAC 2008 (notamment l'usage des fragments — *snippets*), d'autres facteurs pouvaient jouer un rôle positif et qu'une combinaison appropriée de traits pertinents permet d'obtenir de bons résultats.

L'approche expérimentale adoptée permet de préciser l'intérêt des différents facteurs traditionnellement décrits pour le résumé automatique de documents (et inversement, l'étude a permis de montrer le peu d'intérêt de certains critères pertinents dans d'autres contextes). Cette étude ouvre de nouvelles perspectives pour l'intégration de techniques d'analyse d'opinions au sein de systèmes de résumé de texte.

Remerciements

Ce travail a été en partie financé par *INFOM@GIC* du Pôle de Compétitivité *CAP DIGITAL*[7].

Références

ASHER N., BENAMARA F. & MATHIEU Y. Y. (2008). Distilling opinion in discourse : A preliminary study. In *Coling 2008 : Companion volume : Posters*, p. 7–10, Manchester, UK : Coling 2008 Organizing Committee.

[6]Le NIST, dans un fichier fourni avec les résultats, propose les regroupements suivants pour faciliter la lecture des résultats : 1. Contenu (Pyramid) ; 2. Lisibilité (Grammaticality, Non-redundancy, Structure/Coherence and Fluency/Readability) et 3. Pertinence globale (Responsiveness).

[7]`http://www.capdigital.com/`

BOSSARD A., GÉNÉREUX M. & POIBEAU T. (2008). Description of the LIPN Systems at TAC2008 : Summarizing Information and Opinions. In *Proceedings of the Text Analysis Conference*, NIST, Gaithersburg.

BOUDIN F. & MORENO J. M. T. (2007). NEO-CORTEX : A Performant User-Oriented Multi-Document Summarization System. In *Computational Linguistics and Intelligent Text Processing*, Heidelberg : Springer, Lecture Notes in Computer Science.

CARBONELL J. & GOLDSTEIN J. (1998). The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proc. of the 21st annual international ACM SIGIR conference on Research and development in IR*, p. 335–336, Melbourne.

CHÂAR S. L., FERRET O. & FLUHR C. (2004). Filtrage pour la construction de résumés multidocuments guidée par un profil. *Traitement Automatique des Langues*, **45**(1).

CRISPINO G. & COUTO J. (2004). Construction automatique de résumés. Une approche dynamique. *Traitement Automatique des Langues*, **45**(1).

DEY L. & HAQUE S. K. M. (2008). Opinion mining from noisy text data. In *AND '08 : Proceedings of the second workshop on Analytics for noisy unstructured text data*, p. 83–90, New York, NY, USA : ACM.

EDMUNDSON H. P. (1969). New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, **16**(2).

FARZINDAR A., LAPALME G. & DESCLÉS J.-P. (2004). Résumé de textes juridiques par identification de leur structure thématique. *Traitement Automatique des Langues*, **45**(1).

GOTTI F., LAPALME G., NERIMA L. & WEHRLI E. (2007). GOFAIsum : A Symbolic Summarizer for DUC. In *Document Understanding Conference*, Rochester.

GOULET M.-J. (2007). Terminologie et paramètres expérimentaux pour l'évaluation des résumés automatiques. *Traitement Automatique des Langues*, **48**(1).

MANI I. (2004). Narrative Summarization. *Traitement Automatique des Langues*, **45**(1).

NENKOVA A., PASSONNEAU R. & MCKEOWN K. (2007). The pyramid method : incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, **4**(2), 1–23.

RADEV D. R., BLAIR-GOLDENSOHN S. & ZHANG Z. (2001). Experiments in single and multidocument summarization using mead. In *First Document Understanding Conference*.

TURNEY P. D. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. In *40th Annual Meeting of the ACL, Philadelphia*.

WILSON T., WIEBE J. & HWA R. (2004). Just how mad are you ? Finding strong and weak opinion clauses. In *Proceedings of AAAI-04, 21st Conference of the American Association for Artificial Intelligence*, p. 761–769, San Jose, US : AAAI Press / The MIT Press.

YOUSFI-MONOD M. (2007). *Compression automatique ou semi-automatique de textes par élagage des constituants effaçables : une approche interactive et indépendante des corpus*. PhD thesis, Université Montpellier II.

YU H. & HATZIVASSILOGLOU V. (2003). Towards answering opinion questions : Separating facts from opinions and identifying the polarity of opinion sentences. In M. COLLINS & M. STEEDMAN, Eds., *Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing*, p. 129–136, Sapporo, JP.

ZHANG W., YU C. & MENG W. (2007). Opinion retrieval from blogs. In *CIKM '07 : Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, p. 831–840, New York, NY, USA : ACM.